MACE2K - Molecular And Clinical Extraction: A Natural Language Processing Tool for Personalized Medicine

GEORGETOWN UNIVERSITY

PI: MADHAVAN, SUBHA

Grant Number: 1 U01 HG008390-01

DESCRIPTION (provided by applicant): The velocity, variety, volume and veracity of data from relevant information sources make it extremely challenging for oncologists to collect and review pertinent data that can support routine personalized treatment for their patients. There is an urgent need to develop data wrangling approaches including Natural Language Processing and information retrieval methods to extract and curate personalized-therapy related publications and clinical trials. Once curated, the structured data can be used by biomedical researchers to generate novel scientific hypotheses, design new studies, obtain a better understanding of biological mechanisms of disease, perform meta-analyses, and create clinical decision support systems. There is an urgent need to develop improved search interfaces specific to the field of personalized therapy, including ways to display, rank, and save results by end users. While several database and web-based keyword search engine algorithms exist, there is a lack of tools that meet the unique challenges of personalized medicine. There is also an urgent need to develop software that allows for verification and validation of information extracted and ranked through computational methods using subject matter expertise to improve the gold standard corpus that can be used for biomedical research into personalized therapies. To address these issues, we will build an innovative software stack (MACE2K) to adapt and extend widely tested Biocreative natural language processing (NLP) tools to automatically retrieve and pre-process targeted therapy information from clinicaltrials.gov, PubMed abstracts as well as open access articles, and conference proceedings. We will build an entity extraction cartridge to accurately parse gene mutations, translocations, gene expression, protein expression, and protein phosphorylation. A marker disambiguation cartridge will be built to assess for trial inclusion or exclusion criteria and to determine marker-related primary endpoints. We will include a ranking cartridge that uses the disambiguated information on markers, drugs and trials to provide a rigorous scoring of trials and studies according to their relevance for personalized medicine. A novel gamification cartridge will be built to allow subject matter experts to verify and validate the information corpus. Our research leverages National Cancer Institute's investments in several programs (many of which we are involved in) including the NCI drug dictionary, National Cancer Informatics Program (NCIP), I-SPY trials, and Center for cancer systems biology (CCSB) to efficiently accomplish our aims. PUBLIC HEALTH RELEVANCE PUBLIC HEALTH RELEVANCE: This project will develop new computational methods and software to retrieve targeted molecular and drug therapy information from multiple sources of big data including: clinicaltrials.gov, PubMed abstracts, open access articles, and conference proceedings. The software can be used by biomedical researchers to generate new hypotheses for research on personalized cancer treatment decisions based on enormous volumes of public data already in existence. A novel gamification component will be built to allow subject matter experts to verify and validate the information corpus to enhance accuracy of the software.